The proposed approach demonstrates how multimodal neurodiagnostic analysis can be transformed into a high-performance, reproducible, and telemedicine-compatible computational system.

By combining mathematical signal formalization with cloud-native execution principles, the framework enables scalable and clinically applicable tremor assessment.

**References**

1. Amazon Web Services. (n.d.). Building serverless multi-tier architectures with Amazon API Gateway and AWS Lambda. https://docs.aws.amazon.com/whitepapers/latest/serverless-multi-tier-architectures-api-gateway-lambda/
2. Біщак Д.С., & Петрик М.Р. (2025). Кореляція між графомоторною активністю та ЕЕГ у пацієнтів із тремором при хворобі Паркінсона. Вчені Записки Таврійського Національного Університету Імені В.І.Вернадського, Серія «Технічні Науки», 2(4), 25–30. https://doi.org/10.32782/2663-5941/2025.4.2/04
3. Bishchak, D., & Petryk, M. (2025). Algorithmic approach to tremor classification based on EEG and graphomotor signals. Scientific Journal of the Ternopil National Technical University, 119(3), 35–44. https://doi.org/10.33108/visnyk_tntu2025.03.035
4. Amazon Web Services. (n.d.). AWS Lambda. https://aws.amazon.com/lambda/
5. Amazon Web Services. (n.d.). Amazon API Gateway. https://aws.amazon.com/api-gateway/
6. Ramírez, S. (n.d.). FastAPI: Modern, fast (high-performance) web framework for building APIs with Python. FastAPI. https://fastapi.tiangolo.com/

**DOI 10.70286/EOSS-16.02.2026.007.86-88**

# CONTROLLED BENCHMARKS FOR STREAMING DECISION SUPPORT IN MULTIMODAL TIME SERIES

**Uzun Illia**
Postgraduate student
Institute of Artificial Intelligence and Robotics
National University «Odessa Polytechnic», Ukraine

**Abstract.** Validating streaming decision support systems under non-stationary conditions requires distinguishing between genuine concept drift (change in data distribution) and modality degradation (sensor faults, noise, missingness). This work presents a methodology for constructing controlled streaming benchmarks with deterministic injections of drift, degradation, and anomalies. We define a generative protocol where concept drift is modeled as a switch in modality relevance, while

degradations are injected as segment-based missingness, noise, or scale shifts. The proposed prequential evaluation protocol ensures causal interpretability of performance metrics and supports the validation of reliability-adaptive pipelines. The benchmark design prioritizes reproducibility by fixing injection scenarios and providing a minimal set of artifacts for tracing "scenario → metric → conclusion".

**Keywords:** machine learning; data analysis; information systems; decision support; time series; multimodal data; concept drift; data quality; streaming evaluation; controlled benchmarks; reproducibility.

In the domain of streaming multimodal decision support, standard evaluation on real datasets often suffers from the lack of ground truth regarding the cause of performance drops [1–3]. A decrease in forecasting accuracy can stem from a change in the underlying process (concept drift) or from a temporary failure of a data source (modality degradation). To rigorously validate adaptive methods, we propose a controlled benchmarking protocol that disentangles these factors through deterministic injection.

**Generative Model and Drift Protocol.** The backbone of the benchmark is a multimodal stream where the target $y_t$ is a time-varying combination of modality-specific latent processes. We model concept drift not as a change in feature distributions (which mimics degradation), but as a relevance switch. The target is generated as:

$$y_{t+1} = \alpha_t x_{t,0}^{(1)} + (1 - \alpha_t) x_{t,0}^{(2)} + \varepsilon_t,$$

where $\alpha_t$ controls the mixing. Drift events are defined as abrupt or gradual changes in $\alpha_t$ (e.g., from 0.9 to 0.1), forcing the optimal predictor to switch reliance from Modality 1 to Modality 2. This formulation allows testing whether an adaptive system allows "relevance drift" while being robust to "feature degradation".

**Degradation Injection Protocol.** Degradations are injected into the test segment of the stream ($T = 8000$, train $< 5000$) using a segment-based approach. We define three canonical stress types: (i) Missingness: random NaN injection with intensity $p \in [0.2,0.9]$; (ii) Noise: additive Gaussian noise $\mathcal{N}(0, \sigma_{noise}^2)$; (iii) Scale Shift: multiplicative scaling by factor $a \sim \mathcal{U}[1-\delta, 1+\delta]$. To validate robustness, we employ an alternating degradation scenario: Modality 1 degrades in segments $[5000,6000)$ and $[7000,8000)$, while Modality 2 degrades in $[6000,7000)$. This forces the system to demonstrate dynamic switching capabilities.

**Anomaly Injection Protocol.** For anomaly detection tasks, we inject discrete events into the target variable $y_t$ on top of the background stream. Types include spikes ($3\sigma \dots 6\sigma$), level shifts (duration 8–24 steps), and variance bursts. Crucially, anomalies are injected independently of modality degradations, creating four confusion quadrants: (clean, normal), (clean, anomaly), (degraded, normal), (degraded, anomaly). This design allows measuring the False Alarm Rate (FAR) specifically on degraded segments.

A schematic of the controlled streaming protocol timeline is shown in Fig. 1.

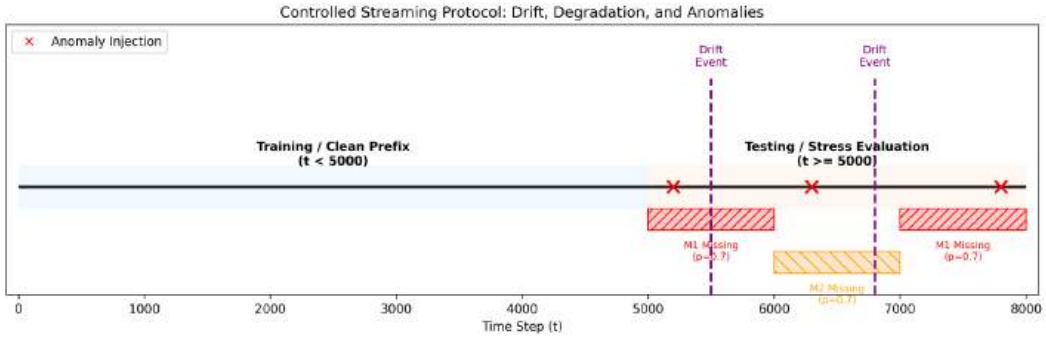


Figure 1. Schematic of the controlled streaming protocol: timeline of drift events, degradation segments, and anomaly injections.

**Evaluation Metrics.** The protocol mandates a prequential (test-then-train) evaluation [6]. For forecasting, we report MAE and RMSE accumulated over specific segments (clean vs. degraded). For adaptation, we measure recovery time (steps to return to baseline error) and update cost (number of samples/parameters). For anomaly detection, we fix a FAR budget (e.g., $\beta = 0.05$) and report Recall@FAR, ensuring that high detection rates are not achieved by flooding the user with false alarms during sensor failures. If probabilistic quality or reliability scores are included in the benchmark, calibration quality can be assessed using standard calibration metrics [4], while threshold selection and separability analysis can be reported via ROC-based tools [5].

The proposed benchmark suite serves as a foundation for validating the reliability-adaptive decision support pipeline (N1–N4). By isolating drift from degradation, it provides the necessary experimental evidence to claim that a system is truly "robust" rather than just insensitive.

## References

1. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
2. Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. Proceedings of the IEEE, 103(9), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697
3. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813
4. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html
5. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
6. Dawid, A. P. (1984). Statistical theory: The prequential approach. Journal of the Royal Statistical Society. Series A (General), 147(2), 278–292. https://doi.org/10.2307/2981683